تعیین خط سیر بهینه در میدان اتومبیل رانی به کمک Monte Carlo

مهران احمدپور

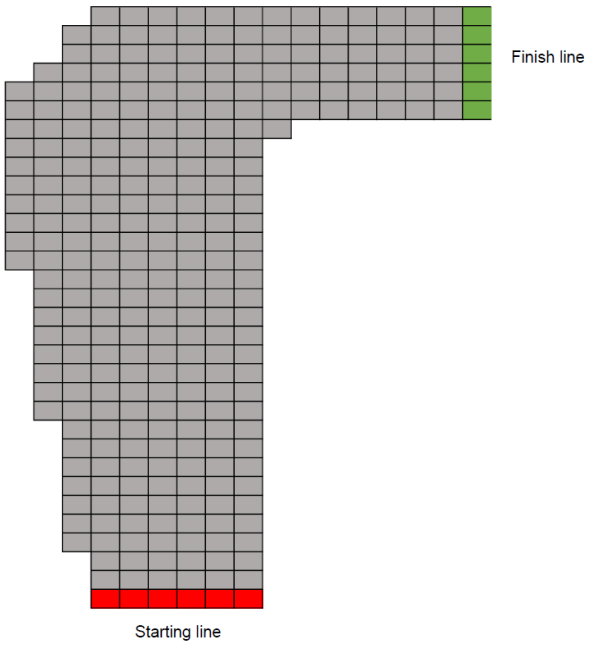
۹۸۲۰۴۹۸۵

چکیده

در این تمرین، با استفاده از روش *Monte Carlo* تلاش شده است تا مسیر بهینه برای یک اتومبیل، در یک میدان، با توجه به قیود سرعت محاسبه گردد. به این منظور پس از طی مراحل شناخت محیط توسط عامل با استفاده از اطلاعات بدست آمده، سیاست تصادفی اتخاذ شده، بهبود یافته و انتخاب اعمال با افزایش مراحل از شکل تصادفی و نرم به شکل حریصانه تغییر داده شده است. در انتها مسیر بهینه برای رسیدن به خط پایان از هر سلول از خط شروع، با توجه به قیود مسئله گزارش شده است.

۱- تعریف مسئله

یک راننده اتومبیل مسابقه‌ای در مسیری پیچ دار از یکی از سلول‌های واقع در خط آغاز تا خط پایان مسابقه می‌دهد. او قصد دارد تا با حداکثر سرعت از پیچ نشان داده شده در شکل-۱ عبور کند ولی به دیواره‌های مسیر برخورد نکند. مسیر مسابقه به صورت یک مجموعه فضای گسسته از سلول ها در نظر گرفته شده است. سرعت حرکت این اتومبیل نیز به شکل دو عدد گسسته و معادل سلول‌های طی شده در هر واحد زمانی در دو مولفه افقی و عمودی است که هر مولفه سرعت می‌تواند دارای تغییرات ۱+ و ۱- و ۰ در هر پله زمانی باشد و هر دو دارای اندازه‌ای کمتر از ۵ می‌باشند.



شکل-۱: محیط فرض شده در مسئله

هدف از طرح این مسئله، محاسبه مسیر بهینه اتومبیل برای رسیدن به خط پایان از هریک از سلول‌های خط آغاز بدون برخورد با دیواره ها و با در نظر گرفتن قیود سرعت اتومبیل می‌باشد.

۲- روش حل و معادلات حاکم

ابتدا در گزارش حاضر الگوریتم مورد استفاده در روش ***Monte Carlo*** شرح داده شده و سپس فرضیات حاکم بر مسئله بیان شده است.

۲-۱) ***On pilicy First visit MC for ε – soft policies***: روش کنترل برخط اولین ملاقات مونت کارلو برای سیاست‌های شبه نرم:

در این روش ابتدا مقادیری دلخواه برای توابع ارزش حالت عمل و سیاست در نظر گرفته می‌شود، سپس عامل در محیط مسئله با اتخاذ عمل‌ها به صورت شبه حریصانه شروع به حرکت کرده و اپیزودها را تشکیل می‌دهد و با ثبت اطلاعات سرعت، موقعیت، اعمال اتخاذ شده و پاداش دریافتی در هر حالت در یک آرایه به یادگیری محیط می‌پردازد. از این اطلاعات در پایان هر اپیزود برای تخمین ارزش اعمال در هر حالت استفاده شده و بهبود سیاست صورت می‌گیرد. به مرور با افزایش تعداد اپیزودها انتخاب اعمال با توجه به سیاست بهبود یافته به شکل حریصانه تری انجام می‌شود. و این فرآیند تا محاسبه مسیر بهینه ادامه خواهد داشت. شکل-۲ الگوریتم ارائه شده برای این روش در مرجع[۱] را نشان می‌دهد.

On-policy first-visit MC control (for ε-soft policies), estimates $\pi \approx \pi_*$

Algorithm parameter: small $\varepsilon > 0$

Initialize:

$\pi \leftarrow$ an arbitrary ε-soft policy

$Q(s,a) \in \mathbb{R}$ (arbitrarily), for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$

$Returns(s,a) \leftarrow$ empty list, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$

Repeat forever (for each episode):

Generate an episode following π: $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$

$G \leftarrow 0$

Loop for each step of episode, $t = T-1, T-2, \ldots, 0$:

$G \leftarrow \gamma G + R_{t+1}$

Unless the pair S_t, A_t appears in $S_0, A_0, S_1, A_1 \ldots, S_{t-1}, A_{t-1}$:

Append G to $Returns(S_t, A_t)$

$Q(S_t, A_t) \leftarrow$ average$(Returns(S_t, A_t))$

$A^* \leftarrow \arg\max_a Q(S_t, a)$ (with ties broken arbitrarily)

For all $a \in \mathcal{A}(S_t)$:

$$\pi(a|S_t) \leftarrow \begin{cases} 1-\varepsilon+\varepsilon/|\mathcal{A}(S_t)| & \text{if } a = A^* \\ \varepsilon/|\mathcal{A}(S_t)| & \text{if } a \neq A^* \end{cases}$$

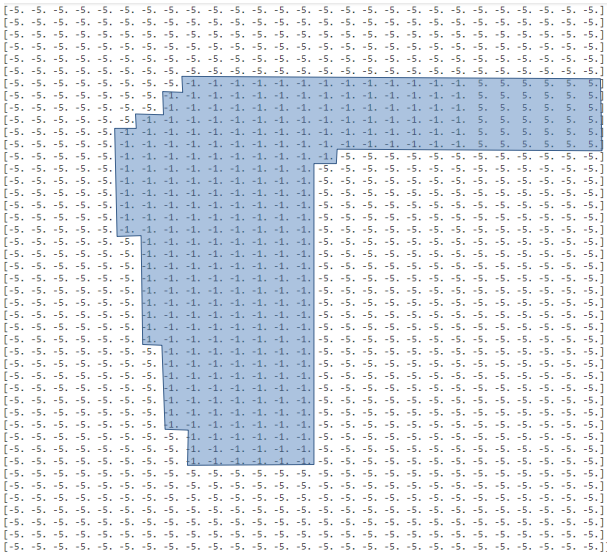
شکل-۲ الگوریتم روش کنترل برخط اولین ملاقات مونت کارلو برای سیاست‌های شبه نرم در مرجع[۱]

۲-۳) فرضیات حاکم بر مساله :

در مسئله مورد بررسی فرضیات زیر لحاظ شده است:

- پاداش هر پله زمانی ۱- ، برخورد با دیوار ۵- و عبور از خط پایان ۵+ فرض شده است.
- پس از برخورد به دیواره‌ها عامل به یک سلول تصادفی در خط شروع باز می‌گردد.
- عامل آزادانه می‌تواند در فضای مسئله به تمامی جهات حرکت کند.
- متغیر ε با افزایش اپیزودها از ۱ به ۰.۱ کاهش می‌یابد.
- پارامتر نرخ تخفیف γ، در مسئله ۰.۹ فرض شده است.
- سرعت عمودی عامل مستقل از تغییرات سرعت افقی آن است و بلعکس.
- ارزش اولیه توابع حالت عمل صفر فرض شده است.
- عامل در هر حالت قادر به تصمیم گیری بین ۹ تصمیم مختلف برای افزایش یا کاهش سرعت های عمودی و افقی خود می‌باشد.

با در نظر گرفتن این فرضیات شکل-۳ پاداش ورود به هر یک از سلول‌های مسئله را نشان می‌دهد.



شکل-۳ فضای در نظرگرفته شده برای حل مسئله

از آنجایی که در مسئله مورد بررسی عامل می‌تواند در هر جهت دلخواهی حرکت کند، همانطور که در شکل-۲ ملاحظه می‌شود فضای مسئله بزرگتر از مسیر تعریف شده در نظر گرفته شده است. این امر به این منظور است که در تمامی حالت‌های خروج عامل از مسیر، سلولی برای موقعیت مقصد تعریف شده باشد. به بیان دیگر اگر عامل با سرعت ۴+ درصدد خروج از دیواره بالایی مسیر باشد سلول‌های ردیف ۲ به عنوان سلول‌های مقصد تعریف شده اند. با توجه به اینکه گام‌های زمانی در این مسئله مورد مطالعه قرار گرفته است، به منظور رفع خطاها در زمان اجرای کد این فضا در نظر گرفته شده است.

اولین سلول در خط آغاز مسیر در ستون ۸ و ردیف ۳۷ فضای مسئله قرار گرفته است و سلول آخر نیز در ستون ۱۳ فضای مسئله تعریف شده است. همچنین خط پایان مسیر در ستون ۲۱ فضای مسئله از ردیف ۶ تا ۱۱ در نظر گرفته شده است.

بدیهیست با تغییر هر یک از فرضیات ذکر شده مسیرهای متفاوتی به عنوان مسیر بهینه محاسبه خواهد شد.

۳- ارائه و تحلیل نتایج

۳-۱) ارائه نتایج

نتایج بدست آمده بعد از طی مراحل یادگیری و بهبود سیاست در این بخش ارائه شده است. باتوجه به اینکه تغییرات سرعت در جهت افقی مستقل از سرعت در جهت عمودی می‌باشد، برای هر سلول مسیرهای متفاوتی می‌توان محاسبه کرد که در پله‌های زمانی یکسانی به خط پایان می‌رسند نتایج ارائه شده تنها برخی از این مسیرها را شامل می‌شود.

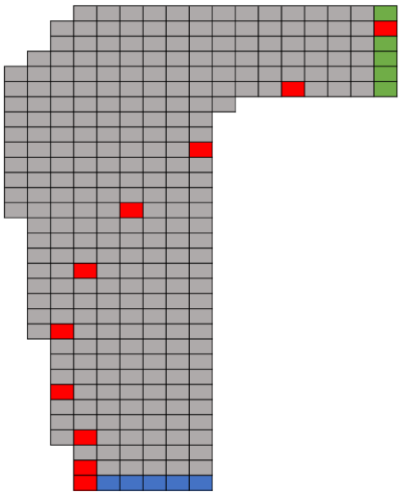
ماتریس نشان داده شده در زیر هریک از مسیرها اطلاعات حالت‌های مختلف عامل را نشان می‌دهد. این اطلاعات از ستون سمت چپ به ترتیب شامل: سرعت در مسیر عمودی، سرعت در مسیر افقی، موقعیت عمودی عامل، موقعیت افقی عامل، شماره تصمیم اتخاذ شده در آن حالت و پاداش ورود به حالت بعد می‌باشد.

در جدول زیر تصمیمات پیش روی عامل در هر حالت بر اساس شماره آنها آورده شده است. عامل در هر حالت می‌تواند بر اساس سیاست تعریف شده و نحوه انتخاب عمل یکی از تصمیمات جدول زیر را اتخاذ نماید.

جدول-۱ شماره تصمیمات پیش روی عامل در هر حالت

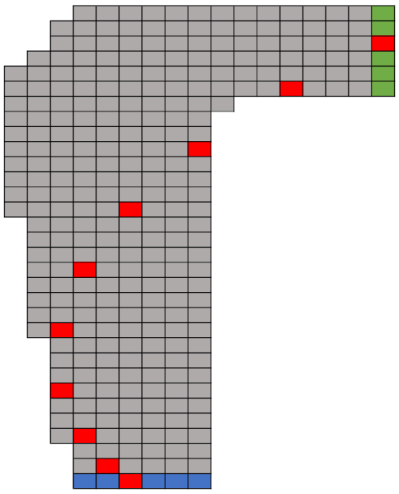
شماره تصمیم	تغییر سرعت عمودی	تغییر سرعت افقی
۰	+۱	۰
۱	۰	+۱
۲	−۱	۰
۳	۰	−۱
۴	+۱	+۱
۵	+۱	−۱
۶	−۱	+۱
۷	−۱	−۱
۸	۰	۰

نتایج محاسبه شده برای هر سلول در خط آغاز در شکل‌های ۴ تا ۹ نشان داده شده است.



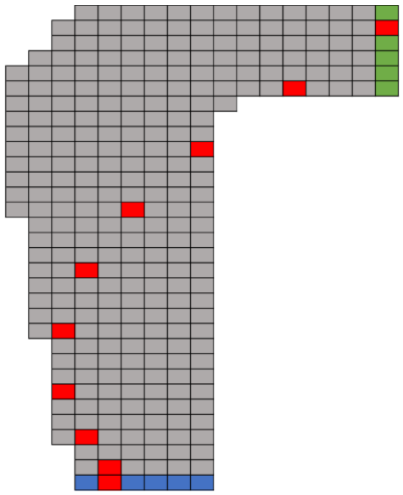
```
[    4     4     7    21     0     0]
[    4     4    11    17     4     5]
[    4     3    15    13     1    -1]
[    4     2    19    10     1    -1]
[    4     1    23     8     4    -1]
[    4     0    27     7     1    -1]
[    3    -1    31     7     4    -1]
[    2     0    34     8     5    -1]
[    1     0    36     8     0    -1]
[    0     0    37     8     0    -1]
```

شکل-۴ مسیر بهینه برای شروع از ستون ۸ خط آغاز



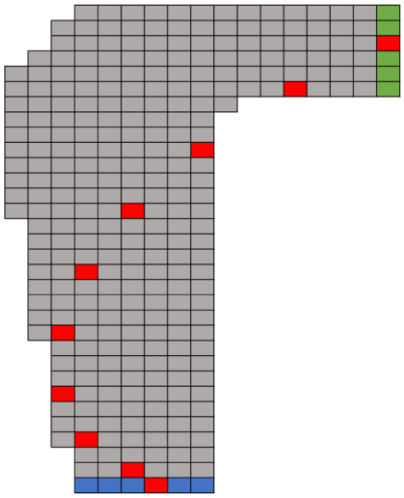
```
[    3     4     8    21     0     0]
[    4     4    11    17     6     5]
[    4     3    15    13     1    -1]
[    4     2    19    10     4    -1]
[    4     1    23     8     1    -1]
[    4     0    27     7     4    -1]
[    3    -1    31     7     4    -1]
[    2    -1    34     8     0    -1]
[    1    -1    36     9     0    -1]
[    0     0    37    10     5    -1]
```

شکل-۶ مسیر بهینه برای شروع از ستون ۱۰ خط آغاز



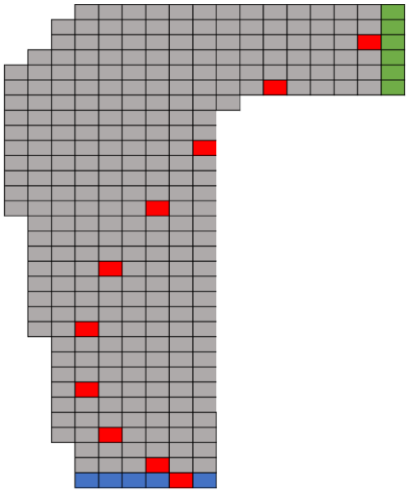
```
[    4     4     7    21     0     0]
[    4     4    11    17     1     5]
[    4     3    15    13     4    -1]
[    4     2    19    10     4    -1]
[    4     1    23     8     1    -1]
[    4     0    27     7     1    -1]
[    3    -1    31     7     4    -1]
[    2    -1    34     8     0    -1]
[    1     0    36     9     5    -1]
[    0     0    37     9     0    -1]
```

شکل-۵ مسیر بهینه برای شروع از ستون ۹ خط آغاز



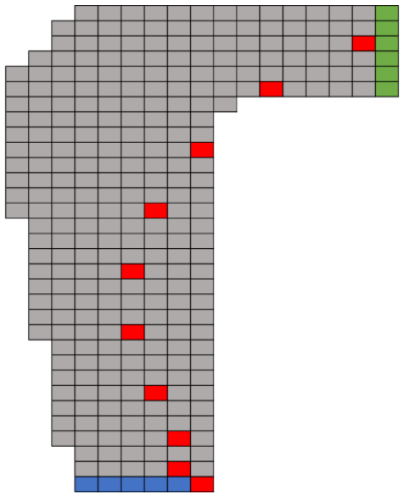
```
[    3     4     8    21     0     0]
[    4     4    11    17     2     5]
[    4     3    15    13     4    -1]
[    4     2    19    10     1    -1]
[    4     1    23     8     1    -1]
[    4     0    27     7     4    -1]
[    3    -1    31     7     4    -1]
[    2    -2    34     8     4    -1]
[    1    -1    36    10     5    -1]
[    0     0    37    11     5    -1]
```

شکل-۷ مسیر بهینه برای شروع از ستون ۱۱ خط آغاز



```
[    2     4     6    24     0     0]
[    3     4     8    20     2     5]
[    4     3    11    16     6    -1]
[    4     2    15    13     4    -1]
[    4     2    19    11     0    -1]
[    4     1    23     9     1    -1]
[    4     0    27     8     1    -1]
[    3    -1    31     8     4    -1]
[    2    -2    34     9     4    -1]
[    1    -1    36    11     5    -1]
[    0     0    37    12     5    -1]
```

شکل-۸ مسیر بهینه برای شروع از ستون ۱۲ خط آغاز



```
[    2     4     6    24     0     0]
[    3     4     8    20     6     5]
[    4     3    11    16     6    -1]
[    4     2    15    13     4    -1]
[    4     1    19    11     1    -1]
[    4     0    23    10     4    -1]
[    4    -1    27    10     1    -1]
[    3    -1    31    11     0    -1]
[    2     0    34    12     5    -1]
[    1    -1    36    12     4    -1]
[    0     0    37    13     5    -1]
```

شکل-۹ مسیر بهینه برای شروع از ستون ۱۳ خط آغاز

۳-۲) تحلیل نتایج:

در نظر گرفتن پله‌های زمانی برای تحلیل حرکت عامل در مسئله مورد مطالعه موجب می‌شود که عامل بدون دریافت پاداش منفی از پیست خارج شده و به آن بازگردد، به نظر می‌رسد مبنا قرار دادن پله‌های مکانی با تغییر در پاداش‌های دریافتی در هر حالت به شکل تابعی از سرعت نتایج دقیق‌تر و منطقی تری را به همراه داشته باشد.

علاوه بر آن فرض مستقل بودن سرعت‌های افقی و عمودی از یکدیگر موجب شده است تا زمان مورد نیاز برای طی یک مسافت مشخص توسط عامل به دو صورت مستقیم و زیگزاگ برابر باشد، در اینجا نیز به نظر می‌رسد فرض محدود بودن مجموع سرعت‌های افقی و عمودی نتایج دقیق‌تری را به همراه داشته باشند. در آن صورت در نظر گرفتن حرکت عامل به سمت چپ فرضی حیاتی برای محاسبه مسیر بهینه خواهد بود. همچنین در صورت باریک بودن مسیر بعد از پیچ مسیر بهینه در صورت در نظر گرفتن حرکت عامل به سمت چپ قابل محاسبه خواهد بود.

در الگوریتم حل مسئله که در فصل-۲ توضیح داده شده است بخش بهبود سیاست نقش بسیار مهمی در تعیین زمان مورد نیاز برای حل مسئله دارد، بعد از اعمال هر بهبود در سیاست می‌بایست روش انتخاب اعمال را حریصانه‌تر نمود تا از بهبود حاصل شده در سیاست برای کاهش مسیر اپیزودها استفاده نمود اما اگر این سیاست به میزان کافی بهبود نیافته باشد کاهش اپسیلون سبب می‌شود تا زمان مسیر هر اپیزود به شدت افزایش پیدا کند، چرا که عامل به انجام یک سیاست غیر بهینه اصرار می‌ورزد. در این راستا روش تخمین ارزش حالت عمل و میزان نرخ تخفیف اهمیت زیادی پیدا می‌کند، با توجه به اینکه مشاهدات نشان می‌دهند که مسیرهای بهینه محاسبه شده در انتهای الگوریتم در همان اپیزودهای اولیه در هنگام یادگیری محیط و قبل از بهبود سیاست به شکل تصادفی توسط عامل انتخاب می‌شوند، انتظار می‌رود بعد از بهبود سیاست در اولین مراحل پاسخ بهینه به سرعت گزارش شود، اما الگوریتم مورد استفاده زمان و اپیزودهای زیادی را برای محاسبه مسیر بهینه صرف می‌کند. به نظر می‌رسد تغییر در نحوه محاسبه تابع ارزش‌حالت عمل با کاهش زمان محاسبه مسیر بهینه همراه باشد.

۴- فهرست علائم

π	سیاست
Q	ارزش حالت عمل
R	پاداش
γ	نرخ تخفیف
S	حالت
G	بازگشت
A	عمل
ε	شاخص تعیین کاوش و بهره‌برداری